Data Immersion

Exercise 5.2

8/23/23

Micky Smith

Data Ethics: Data Bias

1. **Carefully read the background and collection plan again. What Types of potential bias exist in your team lead's collection plan? Why was it biased? Please explain your answer. You may also think of biases that go beyond this reading (e.g., cultural bias).**

   I can find 2 different kinds of bias, collection bias and exclusion bias. The collection bias is because the data collected is from a few years ago and is likely to be outdated. It is possible the drug cartels would have changed their laundering methods, so it would render the data available as to old to be prevalent today. The exclusion bias would be because they gathered information from ATM's within only 100 miles from the border, the sample size would need to be increased to avoid missing important activity.

2. **How might these biases distort the results? What could you do to avoid these biases?**

   The collection bias could distort results by providing inefficient information towards how the drug cartel is currently avoiding fraud and choosing to launder money. The best way to avoid this bias would be to collect more up-to-date data to add to the older data, which would allow potential changes in their patterns.

   The exclusion bias could distort results because it may not have a far enough reach for the distance the drug cartel is willing to go to avoid being caught laundering money. A was to avoid this bias is to continue gathering results for each 100 miles, starting 100 miles from the border, then 200 miles, and so on.

3. **If you know that there is bias in the collection method, what could you do to communicate your concerns to your team lead? Please be as specific as possible.**

   The bias would need to be addressed before providing any results for our research, so speaking with the team lead should be done before any analysis takes place. I would want to provide test results proving that the bias exists and provide the results as well as any concerns the bias may cause to the team lead. Afterwards, I would provide alternative suggestions towards gathering more up-to-date information and expanding the parameters of the locations where the data is gathered.

4. **Read through the details of testing. How might the lack of transparency around the experience and training of the investigators allow for bias?**

   Personal bias seems to be the most abundant in the testing. We are unsure of what the model testing consists of. We know that the analyst looked at numerous variables' of 1000 items, which can always be susceptible to human error. When it comes to the analysts, we are unsure of their background, as well as their training level for the task at hand.

5. **Analyze the bar chart showing the scores of individual analysts and see where their scores fall on the distribution curve. If the mean of the scores was 307 and the standard deviation is 166, which score or scores might you eliminate to control for bias? Why?**

   Any data that falls within one standard deviation (plus or minus 166) of the mean (307) would be considered normal distribution. Analyst 10's score of 759 should be considered for being removed. However, before removing this analysts data, we should review why their information differed so much from other analysts. It's possible outlier data can provide meaning in this research, however, if their results are due to human error (lack of training, experience, etc), it should be removed to avoid skewed results.